UNIVERSIDAD AUTONOMA DE MADRID

Escuela Politécnica Superior

18|19

Trabajo fin de grado

Discriminación Algorítmica

Berta Fernández de la Morena

Escuela Politécnica Superior
Universidad Autónoma de Madrid
C\Francisco Tomás y Valiente nº 11

Campus Internacional excelencia UAM CSIC+

www.uam.es

UNIVERSIDAD AUTÓNOMA DE MADRID
ESCUELA POLITÉCNICA SUPERIOR

Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación

TRABAJO FIN DE GRADO

Discriminación Algorítmica

Estudio del sesgo en arquitecturas de aprendizaje profundo

Autor: Berta Fernández de la Morena
Tutor: Aythami Morales Moreno
Ponente: Julián Fiérrez Aguilar

junio 2019

Berta Fernández de la Morena
Discriminación Algorítmica

Berta Fernández de la Morena

IMPRESO EN ESPAÑA – PRINTED IN SPAIN

A mis padres, Antonio y Alicia.

"Cada día sabemos más y entendemos menos"

Albert Einstein

Agradecimientos

Después de un intenso periodo de 4 años, hoy es el día de escribir este apartado de agradecimientos para finalizar mi Trabajo de Fin de Grado y cerrar así esta etapa de mi vida. Debo agradecer a todos los profesores de la Escuela Politécnica Superior, quienes han compartido su conocimiento y su pasión por la materia, para convertirme hoy en egresada. Ha sido un periodo de aprendizaje profundo, no sólo en el ámbito académico, sino también en el personal.

Me gustaría dar las gracias al grupo de investigación BiDA-lab. Definitivamente me habéis brindado todas las herramientas necesarias para completar mi Trabajo de Fin de Grado satisfactoriamente. En particular, le agradezco a mi tutor Aythami Morales el tiempo que me ha dedicado con su valiosa ayuda, y su capacidad para despertar siempre en mi, interés y curiosidad en los temas tratados. Gracias por confiar en mi.

Merecen una mención especial mis compañeros y amigos de la de carrera: Marta Fernández, Andrea González y Manuel Moyano, han hecho estos años más fáciles en cualquier aspecto. Un placer haberos conocido.

Y por supuesto, el agradecimiento más profundo y sentido va hacia mi familia. A mis padres Antonio y Alicia, por su ejemplo, de constancia y trabajo; y a mi hermana Andrea por sus sabios consejos. Han sido un apoyo constante durante todo mi recorrido académico. Gracias de todo corazón por vuestro aprecio y cariño.

Resumen

En la actualidad la tecnología está cada vez más presente en nuestro día a día. No es de extrañar, por lo tanto, que tecnologías como la biometría estén en pleno crecimiento. El campo de la biometría abarca un gran número de áreas como el reconocimiento de voz, el reconocimiento de huellas dactilares o el reconocimiento facial.

En este trabajo se estudia en profundidad las tecnologías de reconocimiento facial, en particular, se estudia el sesgo o los prejuicios humanos que estos algoritmos puedan presentar en sus resultados. Los sistemas de reconocimiento facial, se están utilizando cada vez con más frecuencia, en la toma de decisiones automatizadas. El hecho de no poder disponer de características e información variada y de esta forma, generalizar los datos para entrenar los modelos puede derivar en un modelo sesgado, y como consecuencia, en la toma de decisiones injustas que impactarían directamente en los usuarios.

El objetivo principal de este estudio se centra en una mejor comprensión del rendimiento alcanzado sobre los diferentes grupos demográficos. También se propone una formulación general sobre la discriminación algorítmica con aplicación a la biometría facial. Los experimentos han sido realizados sobre la nueva base de datos generada, DiveFace, compuesta por 24K identidades de seis grupos demográficos diferentes.

En el marco experimental se consideran dos modelos populares de reconocimiento facial: ResNet-50 y VGG Face. Experimentos que han mostrado fuerte discriminación algorítmica sobre los grupos infra-representados en las bases de datos más populares de imágenes faciales. Esta discriminación puede observarse cuantitativamente en grandes diferencias de rendimiento al aplicar esos modelos sobre diferentes grupos demográficos.

Palabras clave

Reconocimiento facial, sesgo, género, etnia, algoritmo, aprendizaje profundo, red convolucional, modelo, discriminación, clasificación, entrenamiento, rendimiento, base de datos, grupo demográfico.

Abstract

Today, technology is more present in our daily life. It is not surprising, that technologies such as biometrics are growing at a high speed. Biometrics cover a large number of areas such as voice recognition, fingerprint recognition and facial recognition.

In this work, we made a depth study about facial recognition technologies, in particular, we studied the bias or human prejudices that these algorithms may present in their results. Facial recognition systems are being used more and more frequently in automated decision making. The fact of not being able to have characteristics and varied information and to generalize the data to train the models can result in a biased model, and as a consequence, in the unfair decision making that would directly impact the users.

The main aim of this study is focused on the performance achieved over different demographic groups. We also propose a general formulation of algorithmic discrimination with application to face biometrics. The experiments are conducted over the new database generated, DiveFace, composed of 24K identities of six different demographic groups.

Two popular face recognition models are considered in the experimental framework: ResNet-50 and VGG Face. We experimentally show strong algorithmic discrimination over the under-represented groups in the most popular facial image databases. This discrimination can be observed quantitatively in large performance differences when applying these models over different demographic groups.

Keywords

Facial recognition, bias, gender, ethnicity, algorithm, deep learning, convolutional network, model, discrimination, classification, training, performance, database, demographic group

Índice

Listas

Lista de ecuaciones

Lista de figuras

Lista de tablas

1

Introducción

Hoy en día el uso de algoritmos en la toma de decisiones está muy introducido en casi todos los aspectos de la sociedad. Decisiones que tradicionalmente eran tomadas por los seres humanos, como por ejemplo, la concesión de un préstamo, son realizadas hoy en día por medio de algoritmos.

Las grandes empresas tecnológicas están compitiendo continuamente por desarrollar modelos de aprendizaje profundo cada vez más precisos. A medida que más clientes hagan uso de estos algoritmos para automatizar juicios y decisiones importantes, la cuestión del sesgo se volverá crucial. Dado que el sesgo puede introducirse fácilmente en los modelos de aprendizaje automático, existe una gran necesidad de desarrollar métodos de aprendizaje automático que garanticen la equidad en la toma de decisiones.

1.1. Qué es la discriminación y cómo se aplica a los algoritmos

Según la Real Academia de la Lengua Española, la discriminación es el trato desigual que se le da a una persona, o colectividad por motivos raciales, religiosos, políticos, de sexo, de edad, de condición física o mental, etc. El derecho a la no discriminación está profundamente establecido en el marco normativo en el que se basa la Unión Europea. Se localiza en el Artículo 21 de la Carta de los Derechos Fundamentales de la Unión Europea, en el Artículo 14 del Convenio Europeo de Derechos Humanos, y en los Artículos 18-25 del Tratado de Funcionamiento de la Unión Europea.

Los modelos de aprendizaje automático, dependen mayoritariamente de los datos recogidos de la sociedad. En particular, el sesgo en los modelos de machine learning se encuentra asociado a la representación desigual de los datos recogidos. El hecho de no poder disponer de características e información variada y así, generalizar los datos para entrenar los modelos puede resultar en un modelo sesgado, y como consecuencia, en la toma de decisiones injusta que impactaría directamente a los usuarios (Figura 1.1). La discriminación algorítmica tiene lugar cuando un individuo o colectivo recibe un tratamiento injusto como consecuencia de la toma de decisiones algorítmica automatizada [1].

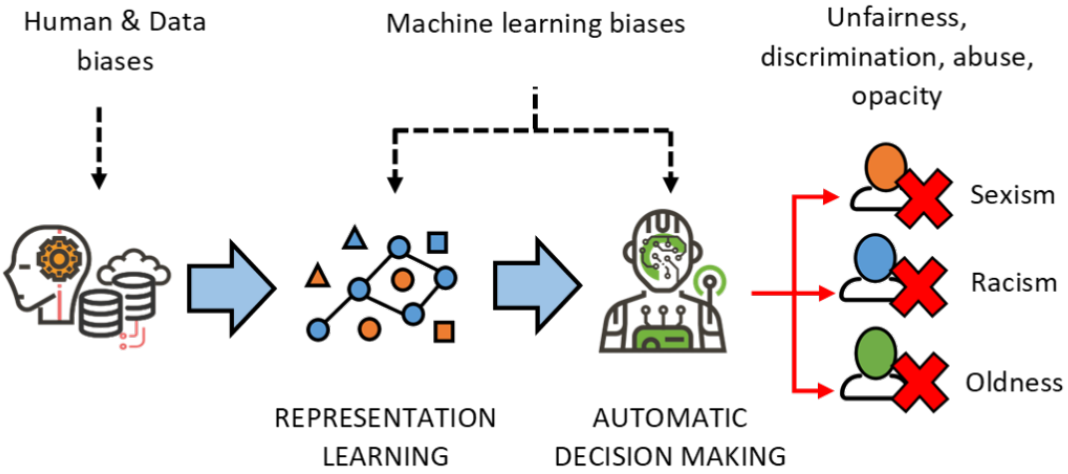


Figura 1.1: Modelo de Aprendizaje Automático. Ilustración extraída de: `https://sensitivenets.com/`

1.2. Sesgo en el Reconocimiento Facial

Podemos afirmar que la inteligencia artificial en el reconocimiento facial presenta un problema destacado, el problema del sesgo. Conceptos como el sesgo y la discriminación, a pesar de no ser conceptos análogos, están muy relacionados. El sesgo se encuentra tradicionalmente asociado a una representación desigual de las clases en una base de datos.

El aprendizaje automático depende en gran medida de los datos recopilados de la sociedad. y en la medida en que la sociedad contenga desigualdad, exclusión u otros rastros de discriminación, también lo harán los datos, reflejando así los sesgos generalizados de la sociedad. En cierto modo, se puede considerar que un algoritmo es tan bueno como la información con la que trabaja. Los datos con frecuencia, son deficientes o inexactos, de manera que se permite que los algoritmos que trabajan con dichos datos hereden los prejuicios de los anteriores responsables de la toma de decisiones [2].

Hoy en día existe un gran número de bases de datos públicas disponibles, compuestas de millones de imágenes faciales que son utilizadas en los modelos de aprendizaje automático. Se ha demostrado recientemente que los algoritmos entrenados con datos sesgados, han dado como resultado algoritmos discriminatorios. Un grupo demográfico subrepresentado en los conjuntos de datos de referencia, puede resultar en discriminación algorítmica [3]. Los sistemas de reconocimiento facial son especialmente sensibles debido a las características faciales del sujeto extraídas de las imágenes, como la identidad, el género, la etnia o la edad.

La tabla 1.1, resume algunas de las bases de datos de imágenes de rostros más populares utilizadas en el reconocimiento facial. Cada una de estas bases de datos se caracteriza por su propio sesgo (por ejemplo, calidad de imagen, pose, fondos y envejecimiento), sin embargo, en este trabajo

destacamos la distribución no uniforme de las clases presente. En todos los casos se considera como clase mayoritaria a los hombres caucásicos. Como se puede observar, las diferencias entre los grupos étnicos son desmesuradas. A pesar de que la población asiática representa alrededor del 65 por ciento de la población mundial, en estas bases de datos analizadas sólo figuran el 9 por ciento de los usuarios.

	#	#	# medio	Asiáticos		Negros		Blancos	
	imag.	ident.	imag./ident.	Muj.	Homb.	Muj.	Homb.	Muj.	Homb.
MS-Celeb-1M [4]	8.5M	100K	85	4.5	7.7	3.9	12.1	19.2	52.4
Megaface [5]	4.7M	660K	7	8.1	10.6	4.7	6.2	30.3	40.0
VGGFace2 [6]	3.3M	9K	370	3.6	3.4	6.3	10.5	30.2	45.9
VGGFace [7]	2.6M	2.6K	1K	2.9	2.1	6.9	5.8	38.6	43.7
YoutubeFaces [8]	621K	1.6K	390	3.0	7.9	4.0	7.7	20.3	56.9
CasiaFace [9]	500K	10.5K	48	2.6	2.6	5.7	7.2	33.2	48.8
CelebA [10]	203K	10.2K	20	5.5	4.4	8.2	6.4	41.5	33.9
PubFig [11]	58K	200	294	1.0	2.0	5.5	6.5	35.5	49.5
IJB-C [12]	21K	3.5K	6	6.2	5.4	6.0	11.8	30.2	40.3
UTKface [13]	24K	-	-	8.9	7.1	16.3	21.5	20.0	26.2
LFW [14]	15K	5.7K	2	2.2	7.2	3.3	9.6	18.7	58.9
BioSecure [15]	2.7K	667	4	4.5	4.3	2.1	3.1	36.0	50.1
Promedio				3.8	4.7	5.3	8.1	28.4	44.5
DiveFace [16]	120K	24k	5	16.7	16.7	16.7	16.7	16.7	16.7

Tabla 1.1: Resumen de la información de algunas de las bases de datos más populares disponibles utilizadas en el reconocimiento facial. De izquierda a derecha: número de imágenes, número de identidades, número medio de imágenes por identidad, y tanto por ciento de imágenes de cada clase (Mujeres,hombres asiáticos; mujeres, hombres negros; mujeres, hombres caucásicos.

Las bases de datos desbalanceadas implican consecuencias negativas para las clases infrarepresentadas. En primer lugar, los modelos son entrenados de acuerdo a una proporción no representativa. Esto provoca que los algoritmos ajusten el modelo a favor de la clase mayoritaria, derivando en métricas de exactitud sesgadas.

Conscientes del problema que esto conlleva, recientemente se han generado bases de datos balanceadas. Estas bases de datos se consideran un recurso valioso, puesto que la amplia diversidad de clases presentes mejoran el rendimiento de la biometría facial. Sin embargo, dado que estas bases de datos no incluyen las identidades de los usuarios, no permiten entrenar los algoritmos de reconocimiento facial, que se basan en la comparación de imágenes de diferentes usuarios para determinar la identidad de los individuos.

Como hemos podido observar, el sesgo es un problema destacado presente en el aprendizaje profundo. En muchos casos, es posible que en el proceso de creación de un nuevo modelo, el sesgo

sea introducido de manera automática. Una vez introducido el sesgo, resulta complicado identificar de donde proviene y como eliminarlo. Adicionalmente, es común diseñar sistemas válidos para diferentes tareas en diferentes contextos. Andrew Selbst, un postdoctorado del Instituto de Investigación de Datos y Sociedad, dice: "Lo que eso hace es ignorar un montón de contexto social, no se puede tener un sistema diseñado en Utah y luego aplicado en Kentucky directamente porque las diferentes comunidades tienen diferentes versiones de justicia. O no se puede tener un sistema que se aplique para obtener resultados 'justos' de justicia penal que luego se apliquen al empleo. La forma en que pensamos sobre la justicia en esos contextos es totalmente diferente".

Por otro lado, muchas de las pruebas estándar realizadas sobre los modelos de aprendizaje profundo no incluyen la detección del sesgo. Dichos modelos son sometidos a pruebas de evaluación de rendimiento anteriores a su implementación, sin embargo, de esta forma no es posible detectar el sesgo, puesto que los datos utilizados para validar el modelo también son sesgados, y no se visualizarán los posibles prejuicios que pueda presentar el modelo [17] .

En los últimos años, el reconocimiento facial ha alcanzado una mayor precisión en diversos escenarios. Esta maximización en el rendimiento ha sido posible gracias a la mejora de los sistemas de cómputo, así como, gracias a la utilización de bases de datos con un número mayor de imágenes. Sin embargo, a la hora de diseñar sistemas biométricos, la precisión que se alcanza como veremos a lo largo de este trabajo, no es el único aspecto a tener en cuenta [18].

1.3. Motivación

Uno de los intereses que ha motivado a la realización de este trabajo ha sido mostrar la naturaleza de ambos conceptos (sesgo de género y etnia) en la forma en la que se manifiestan en los algoritmos de reconocimiento facial.

Dado que hoy en día la toma de decisiones por medio del uso de sistemas automáticos está muy introducido en casi todos los aspectos de la sociedad, existe una gran necesidad de llevar a cabo una evaluación de los sistemas de aprendizaje automático para tomar conciencia sobre la magnitud del problema como consecuencia de un algoritmo discriminatorio.

Actualmente se están tomando medidas al respecto, como por ejemplo, la nueva Regulación Europea de Protección Datos, que fuerza a los responsables del tratamiento de datos a utilizar procedimientos matemáticos o estadísticos con el fin de impedir, entre otros aspectos, efectos discriminatorios por motivo de raza u origen étnicos, religión o creencia, etc. [19].

Además, hay que señalar como dato importante que incluso en el programa electoral de uno de los partidos políticos con más representación en las pasadas elecciones generales de abril de 2019 en España, se incluía la necesidad de eliminar el sesgo de género y otras discriminaciones en el desarrollo

de la Inteligencia Artificial [20]. Esto nos hace pensar que es un tema que preocupa actualmente en todos los ámbitos de la sociedad.

1.4. Objetivos

El objetivo principal de este trabajo es el análisis de los sistemas de reconocimiento facial desde el punto de vista de la discriminación que los mismos puedan presentar. Además, como solución a este problema, una parte importante ha sido la generación de una base de datos diversa, de la que se hablará más adelante.

Otros objetivos a considerar dentro del trabajo son el estudio de diferentes arquitecturas de aprendizaje basadas en redes neuronales profundas (VGG Face y ResNet-50), y su posterior análisis en distintos grupos poblacionales. Además, vamos a observar la posible influencia del sesgo en las etapas de los algoritmos del reconocimiento facial.

Por último, el entrenamiento sobre arquitecturas propias se ha realizado con el objetivo de comprender el funcionamiento o la influencia del sesgo en las mismas.

1.5. Organización

Este Trabajo de Fin de Grado se organiza de la siguiente manera. En primer lugar, se realiza una exposición del estado del arte en el ámbito de la discriminación algorítmica en base al reconocimiento facial. A continuación, se muestra el desarrollo, así como, la integración, pruebas y resultados donde se explica el proceso llevado a cabo, así como los resultados obtenidos. Por último, se concluye el trabajo con una pequeña propuesta de desarrollo futuro.

2

Estado del arte

Estudios recientes han demostrado que los algoritmos de aprendizaje automático basados en clases relacionadas con el género o la raza, pueden llegar a ser discriminatorios, por el hecho de haber sido entrenados con datos sesgados. Además, el hecho de no tener bases de datos diversas, en las que se tenga en cuenta todos los grupos demográficos y distintas características como el color de piel, pelo... conlleva a serias consecuencias en sistemas de visión artificial, por el hecho de tener un peor rendimiento en ciertos grupos poblacionales. En otros contextos, un grupo demográfico infrarepresentado en las bases de batos, puede sufrir frecuentes amenazas. En Estados Unidos, el uso de sistemas de reconocimiento facial está muy introducido en muchos aspectos relacionados con la vigilancia y la prevención de crímenes. Tras una larga investigación se reveló que los individuos afroamericanos eran más propensos a ser detenidos, así como a ser buscados como supuestos sospechosos. Los falsos positivos en estos sistemas, suponen una amenaza para las libertades civiles de los ciudadanos [21]. Además, en *The Perpetual Lineup*", Garvie et al., 2016, se proporciona un análisis profundo de las actividades no reguladas sobre el uso de tecnologías como el reconocimiento facial.

Los algoritmos de detección y clasificación de rostros, también son utilizados en EE.UU. por las fuerzas de seguridad para la vigilancia y prevención de la delincuencia. En otros estudios, se ha demostrado que dichos sistemas de reconocimiento facial presentan una menor precisión en personas etiquetadas como femeninas, negras, o con edades comprendidas entre 18-30 años, frente a la precisión alcanzada para otros grupos demográficos [22].

Nuevamente, en el artículo publicado por Klare et al., 2012, [22] se demuestra un peor rendimiento en tres de los sistemas comerciales de reconocimiento facial más conocidos sobre ciertos grupos poblacionales, en particular, sobre las mujeres, jóvenes (18-30 años) y personas de color. Los resultados obtenidos en los experimentos realizados, permitieron obtener una hipótesis sobre las discrepancias entre algunas de las clases demográficas.

- El reconocimiento facial de mujeres, negros y jóvenes resulta más complicado de realizar.
- El entrenamiento de sistemas de reconocimiento facial sobre conjuntos balanceados y bien distribuidos es imprescindible para reducir la vulnerabilidad sobre determinados grupos sociales.

Cook et al., 2019, ([23]), examinaron el efecto de los factores demográficos en once sistemas biométricos de reconocimiento facial. Dichos sistemas se evaluaron con un proceso repetitivo de medida de eficiencia (tiempo de ejecución) y rendimiento. Además, con el objetivo de observar la influencia del color de piel, desarrollaron un método automático para medir la tonalidad de la piel de los sujetos y cuantificar el efecto que pudiera tener este parámetro u otras covariables demográficas sobre el rendimiento y eficiencia de los sistemas automáticos. Ambas métricas, (eficiencia y rendimiento) en los sistemas evaluados, se vieron afectadas por las covariables demográficas como la piel, género, edad... El color de piel tuvo el efecto más llamativo sobre un modelado lineal, donde se demostró que una piel más oscura estaba asociada con una menor eficiencia (mayor tiempo de transacción) y precisión (peores resultados obtenidos en la similitud).

En *Men also like shopping*, Zhao et al., 2017 ([24]), se realizó un estudio de los datos y de los modelos asociados a la clasificación apoyados por objetos presentes en las imágenes. En dicho estudio tomaron conciencia de que, en primer lugar, las bases de datos contenían grandes sesgos en género, además de que los modelos entrenados con estas bases de datos amplificaban aún más el sesgo existente. Las tareas de reconocimiento visual que involucran el lenguaje, surgieron con la idea de incrementar la cantidad de información que puede adquirirse de las imágenes. Esta tarea tiene como objetivo extraer semántica de las imágenes a partir de grandes cantidades de información etiquetada. Estos métodos suelen combinar predicción estructurada y técnicas de aprendizaje profundo para modelar correlaciones entre las etiquetas y las imágenes para hacer juicios, ya que de otra forma habría un apoyo visual leve. Por ejemplo en la Figura 2.1, es posible predecir una espátula al considerarse una herramienta común en la cocina. Estos métodos corren el peligro de explotar los sesgos que existen en la sociedad, corriendo el riesgo de que magnifiquen estereotipos creados en la sociedad.

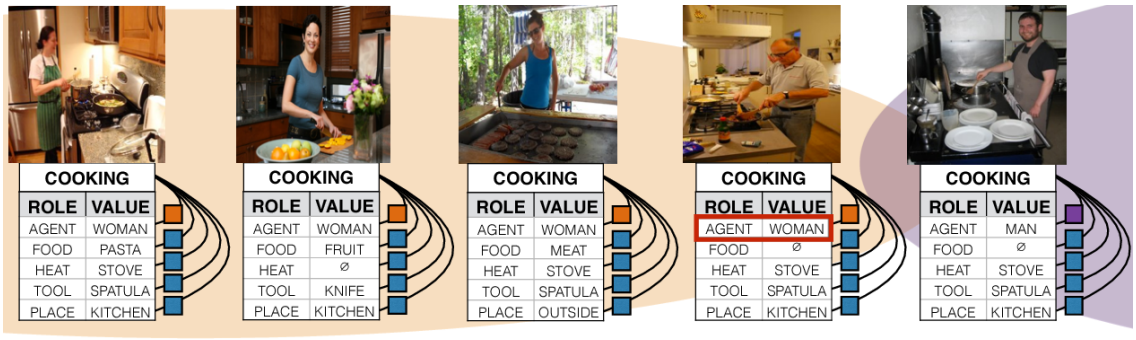


Figura 2.1: Cada imagen está asociada a una tabla con la descripción de la siguiente situación: el verbo,*Cooking*, las funciones semánticas asociadas al verbo, i.e *Agent* así como el valor que tienen en la imagen i.e *woman*. Ilustración extraída de: [24].

3

Desarrollo

3.1. DiveFace

Una de las grandes aportaciones de este Trabajo de Fin de Grado, ha sido la colaboración en la generación de una base de datos diversa (DiveFace) para el entrenamiento de Algoritmos de Reconocimiento Facial y para su posterior evaluación. Esta aportación la he realizado conjuntamente con otra alumna, Marta Fernández de Barrio quien realizaba al mismo tiempo su Trabajo de Fin de Grado.

DiveFace, ha sido generada a partir de la base de datos de entrenamiento Megaface MF2 [25] . Megaface MF2 se trata de una base de datos pública disponible compuesta de 4.7 millones de imágenes faciales con 672 mil identidades diferentes. Las imágenes de cada identidad se encuentran dispuestas dentro de diferentes directorios, lo cual resulta muy útil para entrenar algoritmos para el reconocimiento facial. Las imágenes de Megaface fueron obtenidas de la base de datos Flickr Yahoo´s [26].

DiveFace consiste en una base de datos de imágenes faciales clasificadas dentro de una estructura demográfica. En dicha estructura, los usuarios se encuentran agrupados según su género (masculino o femenino) y sus rasgos físicos étnicos:

- Extremo Oriente. Se considera una región constituida por las regiones de Asia Oriental y Sureste Asiático. Hace referencia a individuos con orígenes ancestrales en China, Japón, Corea, Filipinas, Indonesia y demás países considerados dentro del Extremo Oriente.
- Sub-Sahara, Sur de India. Hace alusión a grupos poblacionales con orígenes ancestrales en el África Subsahariana, India, Bangladesh y Bután entre otros.
- Caucásico. Apunta a personas que tengan sus orígenes en cualquiera de los pueblos originarios de Europa.

En la figura 3.1, se muestra un ejemplo de cada imagen correspondiente a un individuo con origen étnico en uno de los 6 grupos poblacionales presentes en la base de datos DiveFace.

La Base de Datos está compuesta de 24 mil identidades diferentes agrupadas de forma homogénea en estos seis grupos demográficos distintos (cuatro mil identidades para cada clase). El número

(a) Hombre Negro (b) Mujer Negra (c) Hombre Blanco (d) Mujer Blanca

(e) Hombre Asiático (f) Mujer Asiática

Figura 3.1: Ejemplos de imágenes presentes en la base de datos DiveFace.

promedio de imágenes por identidad es de 5.5 imágenes, con un número mínimo de tres imágenes por identidad. En total, hay alrededor de 120 mil imágenes.

Atendiendo a la definición de la R.A.E, la etnia se define como la Comunidad Humana definida por afinidades raciales, lingüísticas, culturares etc.

La clasificación étnica, acorde a la definición dada anteriormente, se ha realizado con el fin de ser exhaustivo, y maximizar las diferencias entre las clases, para que de esta forma se incluyera en alguna de sus categorías a todos los componentes de la población analizada. A pesar de que, durante varios siglos (XVI-XIX), el concepto de raza biológica fue un término incuestionable, así como, el eje central de la antropología, en la actualidad ya no goza de tal aceptación. Hoy en día, los sistemas de clasificación racial están desapareciendo gradualmente de la literatura científica y de los programas de investigación relacionados con la antropología biológica. Un estudio realizado entre especialistas de 13 países, mostró que el 65,7 por ciento de ellos consideró inadecuado el término de raza para hacer referencia a los diferentes grupos de la especie humana. Algunas de las razones para rechazar conceptos como la etnia o la raza son, la generación de posibles conflictos entre grupos sociales, o el impacto en las base de ideologías racistas y xenófobas [27].

Por todo esto, somos conscientes de las limitaciones que existen al clasificar todos los orígenes étnicos humanos en solo 3 categorías.

3.2. Discriminación Algorítmica: Formulación

Con la intención de llevar a cabo un estudio sobre la discriminación en la Inteligencia Artificial en general, se va a presentar una nueva formulación matemática, creada por el grupo BiDA-Lab, basada en la definición anterior dada por la RAE.

A pesar de que se han encontrado ideas similares [28] [29], no se ha enccontrado este tipo de formulación en trabajos similares. Esperamos que la formalización de estos conceptos pueda ser beneficiosa para fomentar la investigación y el debate en este tema candente.

Comencemos con la anotación y las definiciones preliminares. Supongamos que x_s^i es una representación conocida de un individuo i (de entre I individuos diferentes) correspondiente a una muestra de entrada s (de entre S muestras) de ese tema en particular. Se supone que dicha representación x es útil para la tarea T, por ejemplo, para la identificación de caras. Esa representación se aprende utilizando un enfoque de inteligencia artificial con los parámetros θ. También asumimos que hay un criterio de bondad G en esa tarea que maximiza alguna función de rendimiento f dada una base de datos D (conjunto de muestras), es decir, $G(D) = \max_\theta f(D, \theta)$.

Por otro lado, los I individuos pueden clasificarse en función de los criterios demográficos C_d, dados en D, donde $d = 1, , , , D$ (posible fuente de discriminación). Por ejemplo, C_1 = *género = [hombre, mujer]*, en este ejemplo, el criterio demográfico uno es igual a *género* y contiene dos clases: *hombre* y*mujer*. Una clase particular $k = 1, ..., K$, dados un criterio demográfico d y una muestra, se denota como $C_d(x_s^i)$. Por ejemplo $C_1(x_s^i) =$ *hombre*.

Suponemos que todas las clases están uniformemente representadas en el conjunto de datos D, es decir, el número de muestras para cada clase en todos los criterios en D es significativo. $D_d^k \epsilon D$ representa todas las muestras correspondientes a la clase k de un criterio demográfico d.

Finalmente, definimos la Discriminación Algorítmica como el algoritmo que discrimina a un grupo de individuos representados por una clase k (v.g, *mujeres*), cuando se realiza una tarea T (v.g, identificación de caras o el reconocimiento de emociones) si en esa tarea la bondad G al considerar el conjunto completo de datos D (incluyendo múltiples representaciones de múltiples individuos), es significativamente mayor que la bondad $G(D_d^k)$ en un subconjunto de datos correspondiente a la clase k del criterio demográfico d. Nótese que la formulación anterior puede extenderse fácilmente al caso de un número variable de muestras S_i para diferentes sujetos, lo cual es un caso habitual; o a las clases K que no son disjuntas.

Cabe destacar también que la formulación anterior se basa en el rendimiento medio de los grupos de individuos. En muchas tareas de inteligencia artificial es habitual un rendimiento diferente entre individuos específicos debido a diversas razones (por ejemplo, usuarios específicos que no fueron detectados correctamente [30]), incluso en el caso de algoritmos que, en promedio, pueden funcionar de forma similar para las diferentes clases pueden ser fuente de discriminación.

3.3. Modelos Preentrenados Sesgados

En la última década, se han reducido enormemente las tasas de error de los algoritmos de reconocimiento facial gracias a las Redes Neuronales Convolucionales [31]. Estas redes son capaces de aprender características muy discriminatorias a partir de grandes bases de datos. Una alternativa muy común y efectiva para el aprendizaje profundo en conjuntos de datos de imágenes pequeños, es aprovechar una red pre-entrenada. Una red pre-entrenada, es una red guardada, que ha sido previamente entrenada con un gran conjunto de datos, normalmente en una tarea de clasificación de imágenes a gran escala. Si dicho conjunto de datos de entrada es lo suficientemente amplio, entonces, por lo general, la jerarquía de características aprendida por la red pre-entrenada podrá actuar como un modelo genérico para prácticamente cualquier problema de clasificación, aunque estos nuevos problemas, involucren clases completamente diferentes a las de la tarea original.

En los experimentos realizados se parte de dos modelos pre-entrenados para el reconocimiento facial: VGG-Face y ResNet-50. Estos modelos han sido evaluados demostrando un gran desempeño en puntos de referencia públicos de reconocimiento facial [6] [7]. VGG Face es un modelo basado en la arquitectura VGG-Very-Deep-16 CNN (*Convolutional Neuronal Network*), entrenada con el conjunto de datos VGGFace [7]. ResNet-50 es un modelo de Redes Neuronales Convolucionales entrenado con más de un millón de imágenes de la base de datos ImageNet (`http://www.image-net.org`). Esta red está determinada por 50 capas y 41 millones de parámetros propuestos inicialmente para la tareas de reconocimiento de imágenes, siendo capaz de clasificar las imágenes en mil clases de objetos diferentes [32]. La diferencia principal, con respecto a las Redes Neuronales Convolucionales tradicionales, es la incorporación de conexiones residuales entre capas no consecutivas que permiten que la información salte algunas capas mejorando el proceso de aprendizaje. El modelo Resnet-50 utilizado en nuestros experimentos ha sido entrenado con VGG Face de acuerdo con los criterios especificados en [32]. En relación a los detalles de entrenamiento: tasa de aprendizaje: $\alpha = 10^{-4}$; optimizador: *Adam* con $\beta_1 = 0{,}9$, $\beta_2 = 0{,}999$ y $\epsilon = 10^{-8}$.

Como se ha observado en la Tabla 1.1, las bases de datos empleadas para entrenar estos dos modelos están muy sesgadas, por lo tanto, se espera que los modelos de reconocimiento facial entrenados con dichos conjuntos de datos presenten discriminación algorítmica.

Cada modelo pre-entrenado se utiliza como extractor de vectores de características. La extracción de características consiste en utilizar las representaciones aprendidas por la red pre-entrenada para extraer características relevantes de nuevas muestras de entrada. Dichas características, pasan por un nuevo clasificador que se entrena desde cero (Figura: 3.2).

Antes de aplicar los modelos, resulta imprescindible procesar las imágenes sobre las que se va a aplicar dicho modelo, realizando un escalado, en particular de 224x224 píxeles, a través de una interpolación bilineal.

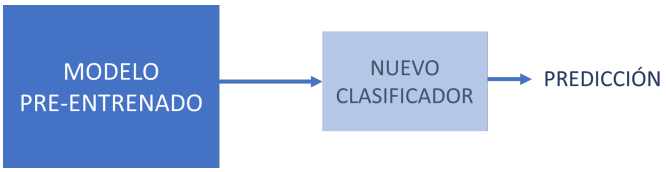


Figura 3.2: Arquitectura de un sistema de clasificación a partir de un modelo pre-entrenado.

3.3.1. Reconocimiento Facial

Uno de los objetivos de este trabajo es evaluar la discriminación algorítmica en la tarea del reconocimiento facial, a partir de los modelos pre-entrenados VGG-Face, ResNet-50. Los vectores de características de cada imagen facial de entrada, se extraen de la última capa de clasificación de cada modelo. El número de características extraídas es, 4096 y 2048 para VGGFace y ResNet-50 respectivamente (Figura: 3.3).

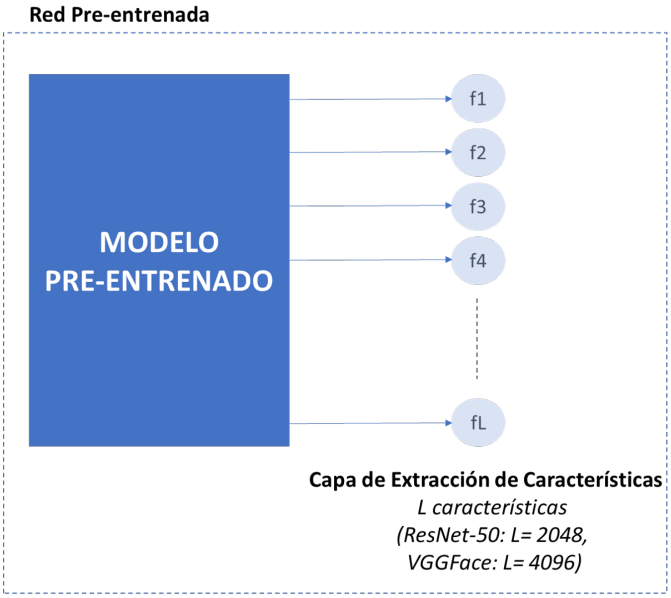


Figura 3.3: Arquitectura de un sistema de identificación de individuos a partir de un modelo pre-entrenado.

Estos vectores de características, una vez extraídos, son normalizados de acuerdo a la norma *l2*, o norma euclídea, para generar la entrada x. La similitud entre dos vectores de características de dos imágenes faciales, x e y, se calcula como la distancia euclídea, $\|x - y\|$. Cabe diferenciar entre distancias genuinas e impostoras. En el primer caso la distancia o similitud se realiza sobre dos imágenes diferentes, $u \neq v$, que pertenecen al mismo usuario i (i.e, $x = {x_{su}}^i$ y $y = {x_{sv}}^i$). La distancia impostora, se calcula sobre de dos imágenes distintas, $u \neq v$, pertenecientes a diferentes identidades, $i \neq j$ (i.e, $x = {x_{su}}^i$ y $y = {x_{sv}}^j$). La tasa de precisión (*Accuracy*), se obtiene mediante la comparación de dichas distancias. Se asignan dos caras a la misma identidad si la distancia obtenida es menor que un umbral, τ.

3.3.2. Clasificación de Género

En esta sección vamos a analizar nuevamente la discriminación algorítmica, pero esta vez, en base al reconocimiento de género. El reconocimiento de género, es un algoritmo de clasificación binaria (Clase positiva: hombre, Clase negativa: mujer), capaz de predecir a que clase va a pertenecer una nueva instancia, basándose en lo aprendido en instancias anteriores (aprendizaje supervisado).

Nuevamente partimos de los modelos pre-entrenados (VGG Face o Resnet-50), a los cuales, se les ha añadido una pila de capas lineal (*Sequential*), además de una capa completamente conectada con 2 unidades de salida (*Dense(2)*), especificando el tamaño esperado de los datos de entrada en función del modelo, 4096 ó 2048 para VGGFace ó Resnet-50 respectivamente. Además, se ha utilizado *Softmax* como función de activación para la capa de salida, compuesta por dos neuronas para clasificar entre mujeres y hombres (Figura: 3.4).

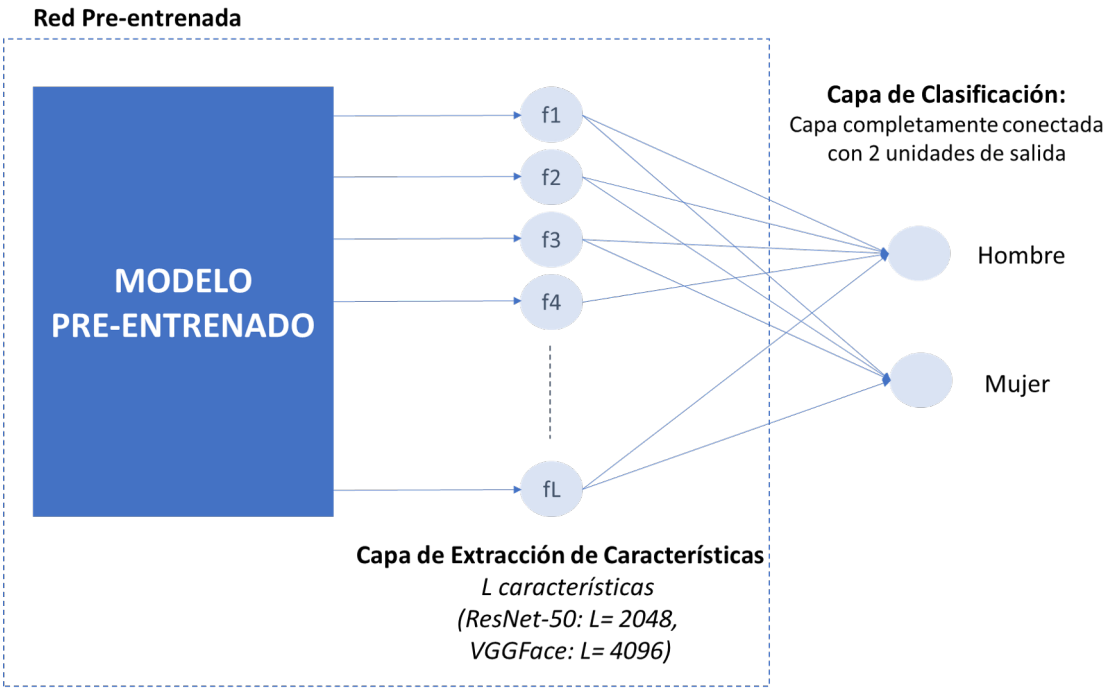


Figura 3.4: Arquitectura de un sistema de clasificación de género a partir de un modelo pre-entrenado.

Previamente al ajuste del modelo, ha sido necesario configurar el proceso de aprendizaje, que se realiza a través del método *Compile*, el cual recibe tres argumentos, un optimizador, una función de pérdidas y una lista de métricas: *Adam*, *Categorical_ crossentropy*, *Accuracy* respectivamente. Por medio de la función *Fit*, se ha ajustado el modelo para un número de épocas (iteraciones en el conjunto de datos). A continuación, para las muestras de entrada se generan las predicciones de salida por medio de la función *Predict*.

3.3.3. Clasificación de Etnia

En esta sección se va a examinar el algoritmo de clasificación de etnia, Se trata de un algoritmo de clasificación multi-clase, capaz de predecir a que clase va a pertenecer una nueva instancia, basándose en lo aprendido en instancias anteriores (aprendizaje supervisado). En particular, se trata de un modelo con tres clases, correspondientes a los tres grupos étnicos presentes en la base de datos DiveFace (Sección 3.1), Blancos, Negros y Asiáticos.

Para la construcción de este sistema de clasificación, se han aprovechado nuevamente los modelos pre-entrenados (VGG Face o Resnet-50), a los cuales, se les ha añadido una pila de capas lineal (*Sequential*), además de una capa completamente conectada con 3 unidades de salida (*Dense(3)*), especificando el tamaño esperado de los datos de entrada en función del modelo, 4096 ó 2048 para VGG Face ó Resnet-50 respectivamente. Además, se ha utilizado *Softmax* como función de activación para la capa de clasificación, la cual, se constituye por tres neuronas para clasificar entre blancos, negros y asiáticos (Figura: 3.5).

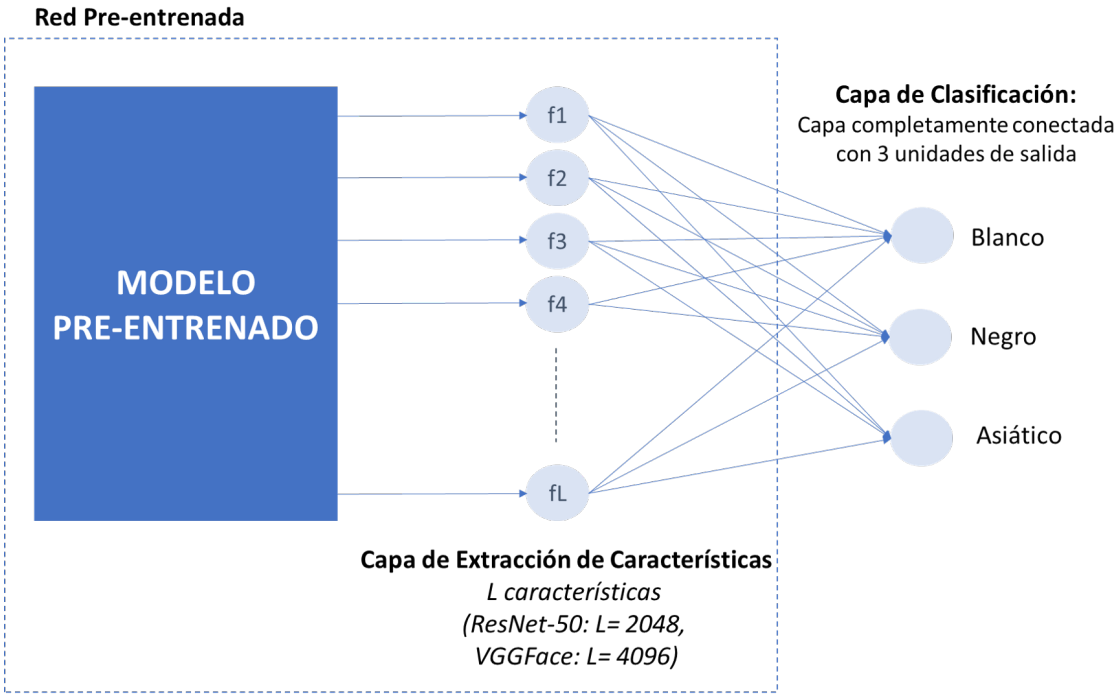


Figura 3.5: Arquitectura de un sistema de clasificación de etnia a partir de un modelo pre-entrenado.

Para llevar a cabo el entrenamiento del nuevo modelo, ha sido necesario configurar previamente el proceso de aprendizaje por medio del método *compile*. Este método recibe tres argumentos, un optimizador, una función de pérdidas y una lista de métricas, en particular, *Adam*, *Categorical_ crossentropy*, *Accuracy* respectivamente. Por medio de la función *Fit*, se ha entrenado el modelo para un número de determinado de épocas (iteraciones en el conjunto de datos). A continuación, la función *Predict* ha permitido generar las predicciones de salida a partir de las muestras de entrada.

3.4. Entrenamiento de una red Convolucional desde cero

En este apartado vamos a construir desde cero una arquitectura de Redes Convolucionales, también conocidas como *ConvNets*. Una de las principales características del aprendizaje profundo es su capacidad de encontrar características discriminantes en los datos de entrenamiento, lo que es posible cuando las muestras de entrada tienen grandes dimensiones, como es el caso de las imágenes. Las Redes Convolucionales son muy eficientes en tareas perceptivas debido a que aprenden características locales e invariantes.

Como ejemplos prácticos, me voy a centrar en la clasificación de las imágenes faciales según su género y su etnia. El problema, por tanto, va consistir en entrenar un nuevo modelo desde cero sobre la base de datos de DiveFace (Sección: 3.1). El entrenamiento desde cero sobre dicho conjunto de datos producirá resultados razonables, sin la necesidad de utilizar otro tipo de extractores de características personalizados para la tarea.

La arquitectura de la Red Convolucional construida se ilustra en la figura 3.6, se compone de una pila de capas *Conv2D* alternadas con capas *MaxPooling2D*, ambas con activación *Relu*. Adicionalmente, se añade una capa *Flatten* que tiene como función convertir la matriz procedente de la capa anterior en un vector "plano". Posteriormente se añaden dos capas ocultas, junto con una capa intermedia *DropOut* la cual elimina al azar algunas de las unidades de cada capa. La primera capa completamente conectada estará compuesta por 512 nodos, mientras que la de salida tendrá 2-3 nodos, en función de la tarea de clasificación (género, etnia respectivamente), con activaciones *Relu* y *Softmax*.

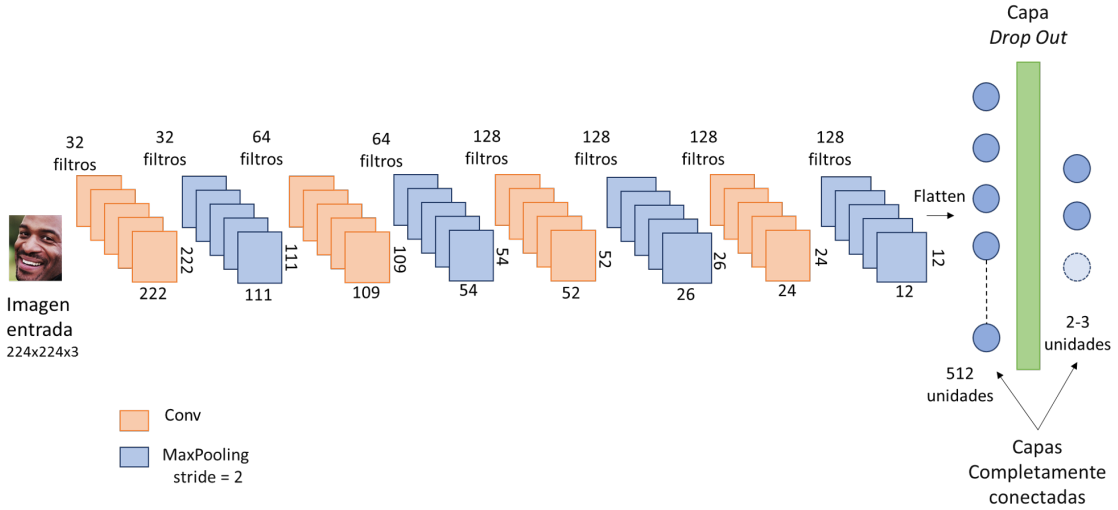


Figura 3.6: Arquitectura de la Red Convolucional construida desde cero.

Una vez definido el modelo, es necesario compilarlo, especificando la función de pérdidas, de optimización y la métricas que emplearemos, en particular, *categorical crossentropy*, *Adam*, *Accuracy* respectivamente, funciones que están pre-implementadas en Keras.

Los datos de entrada, antes de ser introducidos a la red, deben ser procesados. Keras contiene un módulo ubicado en *keras.preprocessing.image* con herramientas que permiten realizar este procesamiento. A continuación, se ha ajustado el modelo a nuestros datos, para que posteriormente sea posible la evaluación del mismo. Esto se ha realizado por medio de la función *fit generator*, especificando ciertos parámetros.

4

Integración, pruebas y resultados

En este capítulo se van a estudiar las diferentes arquitecturas de aprendizaje basadas en redes neuronales profundas (VGG Face y ResNet-50), para evaluar sobre el reconocimiento facial, la identificación de género y etnia, la influencia que pueden tener en su rendimiento, los diferentes grupos poblacionales.

Además, con el objetivo de comprender la influencia del sesgo, se han construido arquitecturas propias para las tareas de detección de género y etnia.

4.1. Modelos Preentrenados Sesgados

4.1.1. Reconocimiento Facial

En este aparado se muestran los efectos que tienen los modelos sesgados (ResNet-50 y VGG Face) sobre el rendimiento de los algoritmos de Reconocimiento Facial.

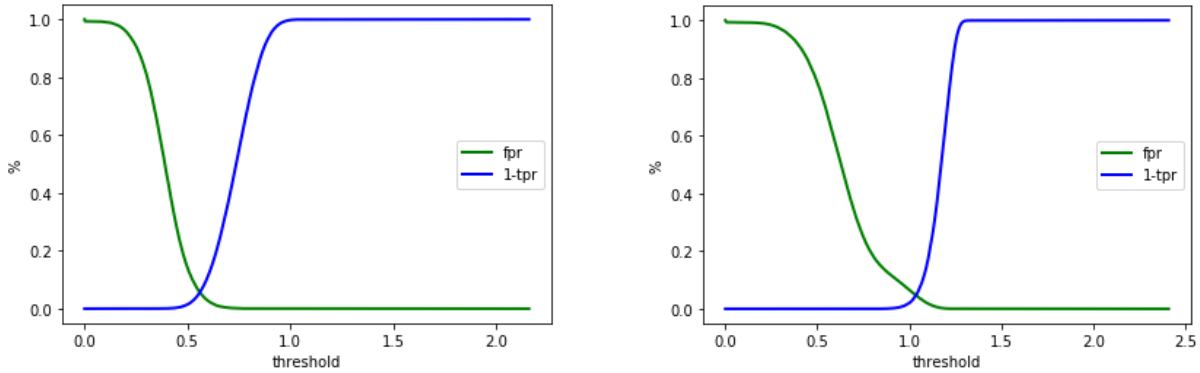
La Tabla 4.1, muestra la exactitud (*ACC, Accuracy*), como medida de rendimiento, para cada grupo demográfico presente en la base de datos DiveFace (sección 3.1). Señalar que cuanto más próximo a 100 % sea este valor, más eficiente es dicho algoritmo.

	Asiáticos		Negros		Blancos	
Modelo	Mujeres	Hombres	Mujeres	Hombres	Mujeres	Hombres
ResNet-50	93.04 %	93.20 %	95.86 %	94.95 %	98.02 %	97.80 %
VGG Face	92.82 %	93.28 %	95.51 %	94.77 %	97.88 %	98.15 %

Tabla 4.1: Exactitud (ACC en %) de la tarea de reconocimiento facial, en base a los modelos preentrenados (VGG Face, ResNet-50) evaluados sobre las clases demográficas presentes en DiveFace (sección 3.1).

Los puntos de referencia que tradicionalmente se toman para evaluar el rendimiento de los algoritmos de reconocimiento facial, no tienen en cuenta estas covariantes demográficas. Los resultados obtenidos en la Tabla 4.1 muestran grandes diferencias en los resultados alcanzados en los diferentes

grupos demográficos, lo que indica que tanto el género como la etnia afectan significativamente en el desempeño de los modelos sesgados. Estos efectos son particularmente altos para la clase menos representada, en particular, para las mujeres asiáticas, presentando una mayor degradación.



(a) VGG (b) ResNet-50

Figura 4.1: Análisis detallado de las curvas FPR(*False Positive Rate*) y 1-TPR = FNR (*False Negative Rate*) en función del umbral sobre el reconocimiento facial.

La figura 4.1, nos aporta información adicional para evaluar las curvas de falsa aceptación y rechazo conjuntamente en función del umbral.

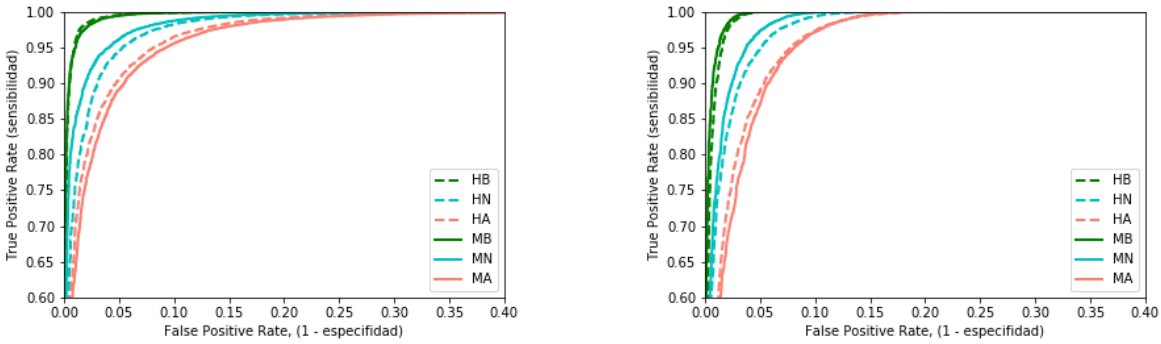
La tasa de falsos positivos, FPR, se define como el cociente entre el número de falsos positivos y el número total de negativos. Hace referencia a las imágenes de diferentes usuarios identificadas incorrectamente como del mismo usuario entre el número de imágenes de diferentes usuarios.

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN} \tag{4.1}$$

Por otro lado, la tasa de falsos negativos hace referencia al cociente entre los falsos negativos y el número total de positivos. Hace referencia a las imágenes del mismo usuario identificadas incorrectamente como de diferentes usuarios entre el número de imágenes del mismo usuario.

$$FNR = \frac{FN}{P} = \frac{FN}{FN + TP} \tag{4.2}$$

Las curvas ROC (*Receiver Operating Characteristic*) que se muestran en la figura 4.2, confirman los resultados heterogéneos. Estas diferencias son importantes ya que marcan el porcentaje de rostros clasificados correctamente, así como, los clasificados incorrectamente. Nuestros resultados suscitan que el origen étnico puede afectar en gran medida las posibilidades de ser clasificado incorrectamente. El bajo rendimiento parece estar originado por una peor capacidad de capturar aquellas características discriminantes para los grupos subrepresentados. Los resultados sugieren que las características

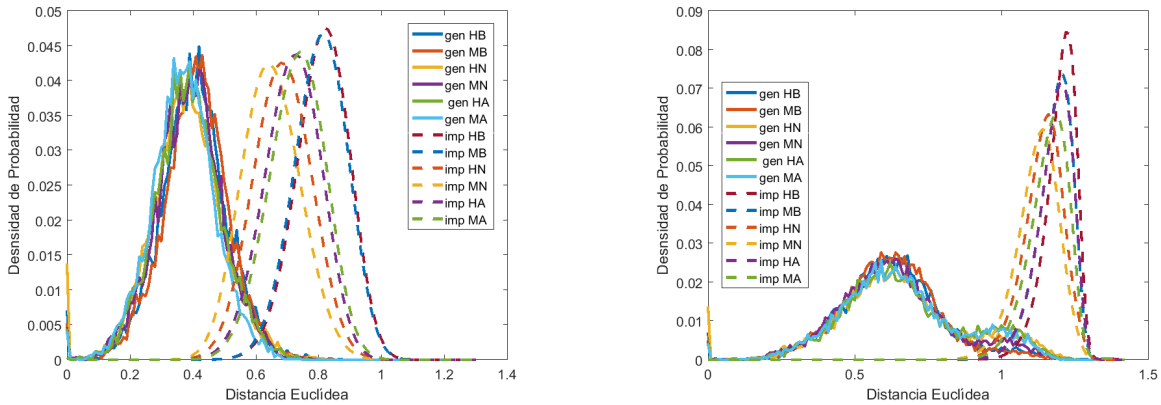


(a) VGG Face (b) ResNet-50

Figura 4.2: Curvas ROC (*Receiver Operating Characteristic*) en la tarea de reconocimiento facial en base a los modelos pre-entrenados VGG Face y ResNet-50.

capaces de alcanzar una alta precisión para una determinado grupo demográfico puede ser menos competitivo en otros.

Vamos a analizar las causas de estas degradaciones. La figura 4.3, representa los histogramas normalizados de las distancias genuinas e impostoras para cada modelo evaluado. En primer lugar, se puede observar claramente como las distribuciones impostoras (variabilidad inter-clase) presentan grandes diferencias, a diferencia de las distribuciones genuinas (variabilidad intra-clase) que se muestran de forma aproximada. Este comportamiento está presente en ambos modelos (4.3(a), 4.3(b)). Además, se puede afirmar que los modelos evaluados presentan dificultades para diferenciar atributos faciales de diferentes sujetos, puesto que la distancia euclídea es mayor.



(a) VGG Face (b) ResNet-50

Figura 4.3: Histograma normalizado de las distancias genuinas (línea continua) e impostoras (línea discontinua) en base a los modelos pre-entrenados VGG Face, ResNet-50 en función del grupo demográfico (H: hombre, M: mujer; B: blanco, N: negro y A: asiático).

4.1.2. Clasificación de Género

En este apartado se va a evaluar el desempeño del algoritmo de detección de género (aprendizaje supervisado), utilizando como herramienta la matriz de confusión.

La tarea de clasificación de género se considera un problema de clasificación binaria, en el cual, se dispone de dos clases, hombre y mujer, consideradas como clase positiva (1) y negativa (0) respectivamente.

En este primer experimento, el modelo se va a entrenar con 2.500 usuarios diferentes para cada uno de los seis grupos demográficos presentes en la base de datos DiveFace (Sección 3.1), es decir, el modelo, se ha va ajustar con un total de 45.000 imágenes (6 grupos demográficos, 2.500 usuarios por grupo demográfico y 3 imágenes por usuario).

		Predichas	
		Mujeres	Hombres
Reales	Mujeres	4469	31
	Hombres	102	4398

(a) ResNet-50

		Predichas	
		Mujeres	Hombres
Reales	Mujeres	4482	18
	Hombres	59	4441

(b) VGG Face

Tabla 4.2: Matrices de confusión obtenidas en la tarea de clasificación de género para cada modelo pre-entrenado(ResNet-50, VGG Face).

Como se puede observar en la tabla 4.2, cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila hace referencia a las instancias en la clase real. Se han definido varios términos estándar para medir el desempeño de un clasificador a partir de la información obtenida de las matrices de confusión, tales como, la tasa de error, la exactitud, la especifidad y la sensibilidad entre otras.

La tabla 4.3, aporta de forma comparativa algunas métricas relevantes para evaluar el algoritmo de la detección de género. De izquierda a derecha se muestra:

- Verdaderos Negativos (*True Negative*): individuos de género femenino identificados correctamente. Mujeres identificadas correctamente como mujeres.
- Falsos Positivos (*False Positive*): individuos de género femenino identificados incorrectamente. Mujeres identificadas incorrectamente como hombres.
- Falsos Negativos (*False Negative*): individuos de género masculino identificados incorrectamente. Hombres identificados incorrectamente como mujeres.
- Verdaderos Positivos (*True Positive*): individuos de género masculino identificados correctamente. Hombres identificados correctamente como hombres.
- Exactitud (*Accuracy*): se define como la proporción del número total de predicciones que

fueron correctas.

- Sensibilidad (*True Positive Rate*): se define como la probabilidad de clasificar correctamente a los individuos con género masculino.
- Especificidad (*True Negative Rate*): se define como la probabilidad de clasificar correctamente a los individuos con género femenino.

	TN	FP	FN	TP	*Accuracy* $\frac{TP+TN}{TN+TP+FN+FP}$	TPR $\frac{TP}{TP+FN}$	TNR $\frac{TN}{TN+FP}$
ResNet-50	4469	31	102	4398	98.52 %	97.73 %	99.31 %
VGG Face	4482	18	59	4441	99.14 %	98.69 %	99.60 %

Tabla 4.3: Metricas para medir el desempeño de la clasificación de género en base a los modelos pre-entrenados ResNet-50 y VGG Face. Información obtenida de las matrices de confusión de la tabla 4.2.

Este modelo de detección de género nos muestra una exactitud del 98.52 % y 99.14 % para los modelos ResNet-50 y VGG Face respectivamente, valores relativamente buenos, donde el modelo VGG Face alcanza una mayor tasa de aciertos en la tarea. Sin embargo, la exactitud no es el único aspecto a tener en cuenta, la sensibilidad o la especifidad resultan también medidas interesantes ya que exhiben un mejor desempeño en la detección de mujeres(clase negativa) que en la de hombres(clase positiva). No obstante, estos resultados dependen del umbral escogido, pudiendo variar las tasas de acierto (TPR, TNR).

A continuación, se va a estudiar la influencia de los rasgos étnicos en la identificación de género, en base a los dos modelos pre-entrenados, ResNet-50(izquierda), VGG Face(derecha). La tabla 4.4, nos muestra una matriz de confusión diferente para cada grupo étnico: 4.4(a),4.4(b) Blancos; 4.4(c),4.4(d) Negros; 4.4(e)4.4(f) Asiáticos.

Para analizar esta información se ha construido nuevamente una tabla con algunas métricas relevantes para evaluar el algoritmo en esta tarea (tabla 4.5). En primer lugar, haciendo una comparativa visual se puede decir que este algoritmo presenta una exactitud mayor, con independencia del modelo, cuando el individuo a clasificar es blanco. Se trata de un efecto muy llamativo, puesto que, se observa que el color de piel ,entre otras características raciales, está asociado a una eficiencia menor.

Las curvas ROC (*Receiver Operating Characteristic*) que se muestran en la figura 4.4, confirman los resultados heterogéneos. Estas curvas aportan el rendimiento de todos los costes de clasificación errónea, el rendimiento de todas las clases, así como una comparativa directa entre los diferentes modelos utilizados (VGG Face: 4.4(a) y ResNet-50: 4.4(b)). En esta figura se encuentran notables diferencias en los resultados obtenidos, lo cual nos sugiere que, el origen étnico puede afectar en gran medida a las posibilidades de ser emparejado incorrectamente. Esto puede deberse a una capacidad menor de capturar las características discriminantes de los grupos subrepresentados de las bases de

		Predichas	
		Mujeres	Hombres
Reales	Mujeres	1497	3
	Hombres	29	1471

(a) Blancos ResNet-50

		Predichas	
		Mujeres	Hombres
Reales	Mujeres	1499	1
	Hombres	22	1478

(b) Blancos VGG Face

		Predichas	
		Mujeres	Hombres
Reales	Mujeres	1486	14
	Hombres	34	1466

(c) Negros ResNet-50

		Predichas	
		Mujeres	Hombres
Reales	Mujeres	1492	8
	Hombres	10	1490

(d) Negros VGG Face

		Predichas	
		Mujeres	Hombres
Reales	Mujeres	1486	14
	Hombres	39	1461

(e) Asiáticos ResNet-50

		Predichas	
		Mujeres	Hombres
Reales	Mujeres	1491	9
	Hombres	27	1473

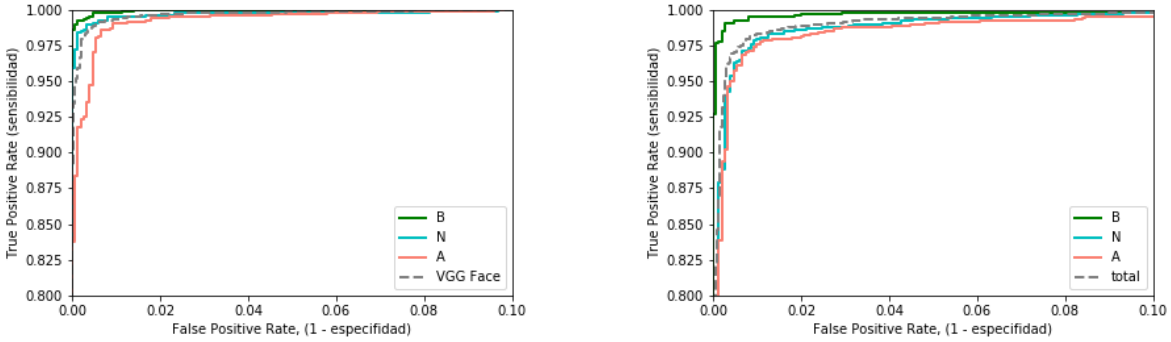
(f) Asiáticos VGG Face

Tabla 4.4: Matrices de confusión obtenidas sobre cada grupo étnico en base a los modelos pre-entrenados (ResNet-50, VGG Face) para la clasificación de género.

		TN	FP	FN	TP	*Accuracy* $\frac{TP+TN}{TN+TP+FN+FP}$	TPR $\frac{TP}{TP+FN}$	TNR $\frac{TN}{TN+FP}$
ResNet-50	Blancos	1497	3	29	1471	98.93 %	98.07 %	99.80 %
	Negros	1486	14	34	1466	98.40 %	97.73 %	99.07 %
	Asiáticos	1486	14	39	1461	98.23 %	97.40 %	99.07 %
VGG Face	Blancos	1499	1	22	1478	99.23 %	98.53 %	99.93 %
	Negros	1492	8	10	1490	99.4 %	99.33 %	99.47 %
	Asiáticos	1491	9	27	1473	98.8 %	98.2 %	99.40 %

Tabla 4.5: Métricas para medir el desempeño de la clasificación de género en base a los modelos pre-entrenados ResNet-50 y VGG Face, en función del grupo étnico. Información obtenida de las matrices de confusión de la tabla 4.4.

datos con las que se entrenó previamente cada modelo, o bien, a una menor capacidad del algoritmo de generalizar las características discriminantes entre hombres/mujeres para diferentes grupos étnicos.



(a) VGG Face

(b) ResNet-50

Figura 4.4: Curvas ROC (*Receiver Operating Characteristic*) en la tarea de clasificación de género en base a los modelos pre-entrenados VGG Face y ResNet-50 en función del grupo étnico (B: Blancos, N: Negros, A: Asiáticos).

Como se ha visto hasta ahora, aprovechar una red pre-entrenada es una técnica muy efectiva para la detección de género, el hecho de entrenar un nuevo clasificador desde cero con un gran conjunto de datos de entrada, amplio y diverso, nos ha permitido alcanzar altas tasas de precisión, tanto en hombres como en mujeres.

Como último experimento, se va a observar la influencia que puede tener en el rendimiento de la tarea, la diversidad de muestras utilizadas para entrenar el nuevo clasificador. Este experimento se realiza en base al modelo ResNet-50.

La Tabla 4.6, muestra la exactitud total, así como la tasa de aciertos en hombres (TPR), y mujeres (TNR), para cada grupo demográfico (Negros, Blancos y Asiáticos), en función de las imágenes o muestras que se utilicen para entrenar el modelo. En particular, los resultados que se muestran en la tabla 4.6(a), se obtienen con 7500 imágenes de mujeres blancas y un intervalo de 7500 a 0 imágenes de hombres blancos para ajustar el modelo. De forma contraria, los datos mostrados en la tabla 4.6(b), se obtienen con 7.500 imágenes de hombres blancos y un intervalo de 7.500 a 0 imágenes de mujeres blancas para ajustar el modelo. Además, la figura 4.5, representa forma gráfica el comportamiento de la tasa de aciertos global (Exactitud en %) presente en la tabla 4.6(a).

En la figura 4.5(a), se observa como a medida que disminuimos el número de muestras con las que ajustamos el modelo, la exactitud en la tarea disminuye, cayendo de forma brusca en 3000 imágenes. Es interesante comentar que cuando se ajusta el modelo únicamente con imágenes de rostros femeninos el algoritmo obtiene el peor resultado, es decir, un *Accuracy* del 50 %. En este caso, el algoritmo es incapaz de clasificar correctamente a los hombres, obteniendo una tasa de aciertos nula (TPR = 0 %). Este comportamiento se observa de manera generalizada, con independencia del grupo étnico al

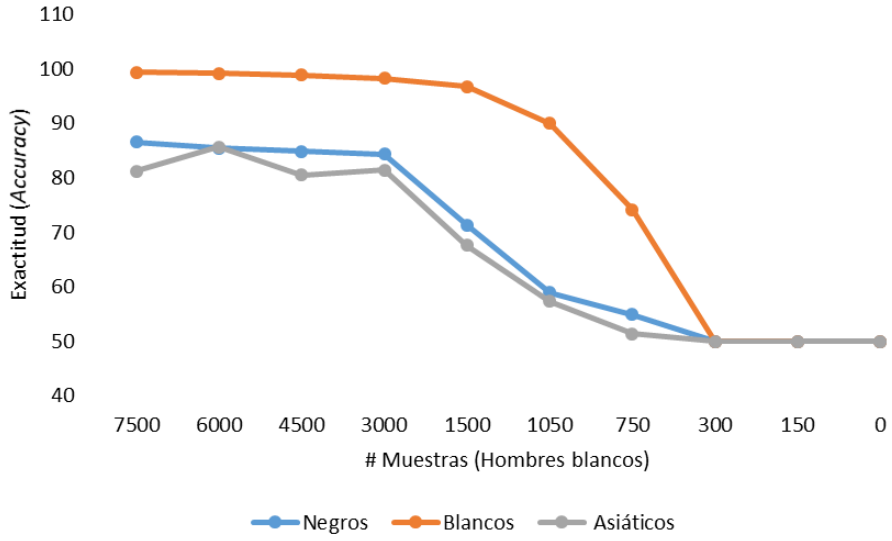
#muestras	Negros			Blancos			Asiáticos		
	ACC %	TPR %	TNR %	ACC %	TPR %	TNR %	ACC %	TPR %	TNR %
7500	86.63	89.09	84.16	99.37	98.79	99.96	81.27	93.72	68.81
6000	85.43	83.92	86.95	99.19	98.41	99.97	85.67	82.99	88.36
4500	84.84	76.97	92.72	98.85	97.73	99.97	80.59	91.64	69.53
3000	84.38	72.55	96.21	98.29	96.61	99.97	81.48	89.79	73.17
1500	71.32	43.02	99.61	96.74	93.49	99.99	67.62	36.64	98.60
1050	58.92	18.09	99.75	89.99	79.99	100.00	57.35	15.72	98.97
750	54.87	9.88	99.87	74.13	48.25	100.00	51.36	2.72	100.00
300	50.00	0.00	100.00	50.00	0.00	100.00	50.00	0.00	100.00
150	50.00	0.00	100.00	50.00	0.00	100.00	50.00	0.00	100.00
0	50.00	0.00	100.00	50.00	0.00	100.00	50.00	0.00	100.00

(a) Hombres.

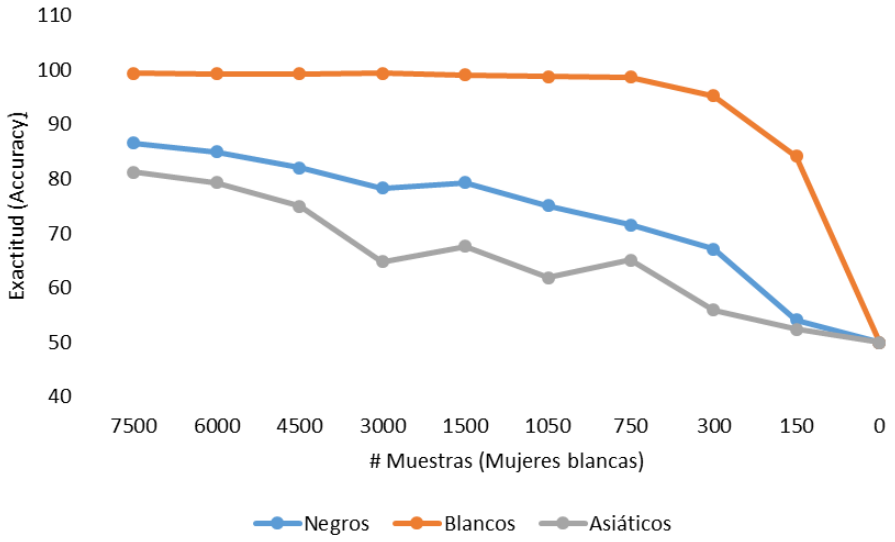
#muestras	Negros			Blancos			Asiáticos		
	ACC %	TPR %	TNR %	ACC %	TPR %	TNR %	ACC %	TPR %	TNR %
7500	86.63	89.09	84.16	99.37	98.79	99.96	81.27	93.72	68.81
6000	84.98	83.95	86.01	99.29	98.67	99.92	79.28	95.01	63.55
4500	82.07	85.80	78.33	99.33	98.87	99.80	75.00	95.11	54.89
3000	78.27	96.29	60.24	99.34	99.24	99.44	64.81	99.15	30.48
1500	79.34	96.44	62.24	99.09	99.36	98.81	67.58	98.83	36.33
1050	75.06	98.32	51.80	98.78	99.64	97.92	61.88	99.47	24.29
750	71.55	98.57	44.52	98.71	99.72	97.71	65.15	99.04	31.27
300	67.1	99.07	35.23	95.30	99.74	90.85	55.93	99.71	12.15
150	54.15	99.87	8.43	84.2	99.92	68.48	52.41	99.96	4.87
0	50.00	100.00	00.00	50.00	100.00	00.00	50.00	100.00	00.00

(b) Mujeres.

Tabla 4.6: Métricas de evaluación del modelo ResNet-50, en la tarea de clasificación de hombres y mujeres blancos, en función del número de muestras de entrada de hombres 4.6(a) y de mujeres 4.6(b).



(a) Hombres



(b) Mujeres

Figura 4.5: Exactitud (*Accuracy* en %) del modelo ResNet-50, en la tarea de clasificación género, en función del número de muestras de entrenamiento de hombres blancos 4.5(a) y de mujeres blancas 4.5(b).

que pertenezca el individuo a clasificar. De igual forma, la figura 4.5(b) representa el comportamiento de la tasa de aciertos global, en función de las muestras de entrada de mujeres blancas con las que el modelo se ajuste. En este caso se observa una caída de rendimiento menos brusca.

Estos resultados exhiben, en primer lugar, la gran importancia que tiene ajustar el modelo con un conjunto de datos amplio, diverso y principalmente balanceado. Como se ha visto, el entrenamiento con conjuntos de muestras desbalanceadas puede suponer una amenaza, en términos de rendimiento, para el grupo subrepresentado. Además, demuestran nuevamente la influencia de las características étnicas sobre la eficiencia de este sistema de clasificación, afectando de forma notable en la probabilidad de ser clasificado incorrectamente.

4.1.3. Clasificación de Etnia

En esta sección se va a evaluar el desempeño del algoritmo de detección de etnia (aprendizaje supervisado). La detección de etnia se trata de un problema de clasificación multi-clase, en particular, tres clases correspondientes a los tres grupos étnicos presentes en la base de datos DiveFace (Sección 3.1: Blancos, Negros y Asiáticos).

En este primer experimento, el nuevo clasificador se va a entrenar con 7500 imágenes de cada uno de los seis grupos demográficos presentes en la base de datos DiveFace (HA,HB,HN,MA,MB,MN), es decir, con un total 45000 imágenes.

		Predichas		
		Blancos	Negros	Asiáticos
Reales	Blancos	2873	56	71
	Negros	16	2943	41
	Asiáticos	11	11	2978

(a) ResNet-50

		Predichas		
		Blancos	Negros	Asiáticos
Reales	Blancos	2962	18	20
	Negros	1	2988	11
	Asiáticos	3	7	2990

(b) VGG Face

Tabla 4.7: Matrices de confusión obtenidas en la tarea de clasifiación de etnia para cada modelo pre-entrenado (ResNet-50, VGG Face)

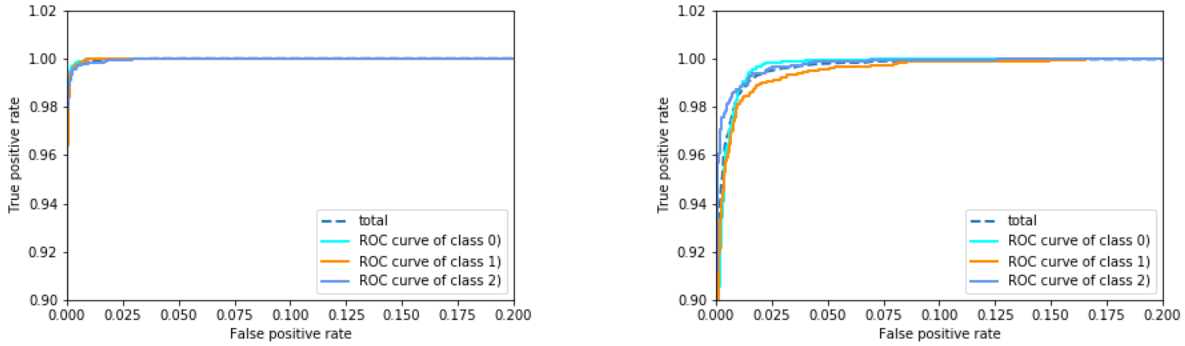
La tabla 4.7 muestra las matrices de confusión como herramienta para evaluar la tarea de clasificación en función del modelo empleado. Para cada matriz de confusión, las columnas representan el número de predicciones de cada clase, mientras que las filas hacen referencia a las instancias en

la clase real. La clasificación de etnia obtiene una exactitud (*Accuracy*) del 97.71 % para el modelo ResNet-50 y del 99.33 % para el modelo VGG Face. Sin embargo, es necesario profundizar un poco más y conocer las predicciones correctas e incorrectas que realiza el clasificador sobre cada clase. En la tabla 4.8, se especifica la tasa de aciertos en cada grupo étnico. Nuevamente, se demuestra que la utilización de una red pre-entrenada es una técnica muy efectiva para tareas de clasificación, en este caso, se han alcanzado tasas de aciertos de más de un 96 %.

	Blancos	Negros	Asiáticos
ResNet-50	95.77 %	98.10 %	99.27 %
VGG Face	98.73 %	99.60 %	99.67 %

Tabla 4.8: Tasa de aciertos en la tarea de clasificación de etnia sobre cada grupo étnico.

La figura 4.6, representa la curva conocida como ROC (*Receiver Operating Characteristic*) para cada modelo pre-entrenado evaluado (ResNet-50, VGG Face). Esta curva aporta el rendimiento de todos los costes de clasificación errónea de cada clase (Clase 0: Blancos, Clase 1: Negros, Clase 2: Asiáticos), así como, una comparativa directa entre los diferentes modelos evaluados 4.6(a) y 4.6(b). Se observa un mayor rendimiento cuando se utiliza como base el modelo pre-entrenado VGG Face.



(a) VGG Face

(b) ResNet-50

Figura 4.6: Curvas ROC (*Receiver Operating Characteristic*) en la tarea de clasificación de etnia (Clase 0: Blancos, Clase 1: Negros, Clase 2: Asiáticos) en base a los modelos pre-entrenados VGG Face y, ResNet-50.

Por último, se va a estudiar en profundidad la influencia del género en la tarea de clasificación étnica. La tabla 4.9, nos muestra una matriz de confusión diferente para cada sexo: 4.9(a),4.9(b) Mujeres; 4.9(c),4.9(d) Hombres.

		Predichas		
		Blancos	Negros	Asiáticos
Reales	Blancos	1443	29	28
	Negros	9	1477	14
	Asiáticos	4	0	1496

(a) Mujeres ResNet-50

		Predichas		
		Blancos	Negros	Asiáticos
Reales	Blancos	1475	13	12
	Negros	0	1494	6
	Asiáticos	1	4	1495

(b) Mujeres VGG Face

		Predichas		
		Blancos	Negros	Asiáticos
Reales	Blancos	1414	32	54
	Negros	0	1480	20
	Asiáticos	6	8	1486

(c) Hombres ResNet-50

		Predichas		
		Blancos	Negros	Asiáticos
Reales	Blancos	1487	5	8
	Negros	1	1494	5
	Asiáticos	1	4	1495

(d) Hombres VGG Face

Tabla 4.9: Matrices de confusión obtenidas sobre cada género en base a los modelos pre-entrenados (Resnet-50, VGG Face) para la clasificación de etnia.

Con la información obtenida de la tabla 4.9, se ha calculado la tasa de aciertos del sistema de clasificación para cada origen demográfico (Blancos, Negros y Asiáticos), en función del género (hombre o mujer) del individuo evaluado y el modelo empleado para la construcción del sistema. Esta información se muestra en la tabla 4.10, cuya interpretación indica que las características faciales discriminantes entre hombres y mujeres, no influyen en la clasificación étnica, es decir, el sistema es capaz de identificar correctamente el origen étnico del individuo con independencia de su sexo.

		Tasa aciertos Blancos (%)	Tasa aciertos Negros (%)	Tasa aciertos Asiáticos (%)	Tasa aciertos global (%)
ResNet-50	Hombres	94.27	98.67	99.07	97.33
	Mujeres	96.20	98.43	99.73	98.13
VGG Face	Hombres	99.13	99.60	99.67	99.47
	Mujeres	98.33	99.60	99.67	99.20

Tabla 4.10: Métricas para medir el desempeño de la clasificación de etnia en base a los modelos pre-entrenados ResNet-50 y VGG Face, en función del género. Información obtenida de las matrices de confusión de la tabla 4.9

4.2. Entrenamiento de una red Convolucional desde cero

4.2.1. Clasificación de Género

En este apartado, vamos a analizar el modelo de clasificación de género sobre una arquitectura de Redes Neuronales construida desde cero. Para ello, he construido la tabla 4.11, que contiene la exactitud del modelo en la detección de género, para cada uno de los experimentos realizados. Este sistema de clasificación ha sido entrenado con diferentes grupos demográficos. De izquierda a derecha.

- Entrenamiento con mujeres y hombres Blancos.
- Entrenamiento con mujeres y hombres Negros.
- Entrenamiento con mujeres y hombres Asiáticos.
- Entrenamiento con mujeres y hombres Blancos, Negros y Asiáticos.
- Entrenamiento con mujeres y hombres Blancos y Negros.
- Entrenamiento con mujeres y hombres Blancos y Asiáticos.
- Entrenamiento con mujeres y hombres Asiáticos y negros.

		Entrenamiento						
		Blancos	Negros	Asiáticos	Blancos Negros Asiáticos	Blancos Negros	Blancos Asiáticos	Asiáticos Negros
Test	Blancos	91.89	87.84	85.05	94.16	94.57	94.57	91.61
	Negros	85.50	91.99	92.91	92.09	89.50	89.50	92.19
	Asiáticos	84.65	82.07	91.37	89.98	93.72	93.85	92.56

Tabla 4.11: Exactitud (*Accuracy* en %) del algoritmo de clasificación de género sobre una arquitectura de redes neuronales construida desde cero.

Estos resultados, nos indican que los individuos con origen étnico blanco, tienen más facilidad de ser emparejados correctamente según su género. Esto puede ser principalmente, porque aquellos rasgos que diferencian entre sexo masculino, y femenino sean mejor identificados en individuos con origen étnico blanco. Aún así, el algoritmo no presenta una gran degradación entre blancos, asiáticos o negros, se puede decir, que es capaz de generalizar con bastante exactitud, los rasgos físicos extraídos de las imágenes faciales que diferencian a las mujeres de los hombres, con independencia del grupo étnico sobre el que se evalúe. Estos resultados, exhiben la necesidad de entrenar los sistemas con conjuntos de datos amplios y balanceados. Tal y como se muestra, el hecho de entrenar un sistema con clases desbalanceadas puede influir en el rendimiento de las clases infrarepresentadas con una mayor probabilidad de ser emparejadas incorrectamente. En términos de exactitud, la degradación puede ser de casi un 10 %.

4.2.2. Clasificación de Etnia

En este apartado, se va a evaluar el algoritmo de clasificación de etnia sobre una arquitectura de Redes Neuronales construida desde cero.

La tabla 4.12, muestra la exactitud (en %) de la tarea de clasificación de etnia, para cada uno de los experimentos realizados, donde en cada uno de ellos se ha entrenado el modelo con diferentes grupos demográficos. De izquierda a derecha.

- Entrenamiento con mujeres Blancas, Negras y Asiáticas.
- Entrenamiento con hombres Blancos, Negros y Asiáticos.
- Entrenamiento con mujeres y hombres Blancos, Negros y Asiáticos.

		Entrenamiento		
		Mujeres	Hombres	Mujeres y Hombres
Test	Mujeres	89.71	82.39	91.38
	Hombres	80.27	90.51	90.42

Tabla 4.12: Exactitud (*Accuracy* en %) del algoritmo de clasificación de etnia sobre una arquitectura de redes neuronales construida desde cero.

Los resultados anteriores exhiben un mejor rendimiento en el reconocimiento de etnia, cuando se evalúa sobre mujeres. Esto puede deberse, a que los rasgos faciales distintivos entre grupos étnicos sean más llamativos en las mujeres que en los hombres. Además, dichos resultados demuestran la importancia de entrenar los modelos con bases de datos diversas y balanceadas, puesto que, como se puede observar, el rendimiento sobre los grupos infrarepresentados puede verse degradado en aproximadamente un 10 %.

5

Conclusiones y Trabajo futuro

5.1. Conclusiones

En este Trabajo de Fin de Grado, se ha intentado presentar un análisis exhaustivo de los modelos de reconocimiento facial desde el punto de vista de la discriminación algorítmica. Hoy en día, las empresas más relevantes en el ámbito tecnológico están compitiendo para ofrecer su modelo de aprendizaje automático, por lo que, resultará crucial que estos modelos garanticen la equidad en el contexto de su aplicación. Como se ha demostrado, los modelos de aprendizaje automático, dependen en gran medida de los datos recopilados para su entrenamiento. Cuando se realiza una clasificación supervisada, como es el caso de los algoritmos de reconocimiento facial o clasificación de género/etnia, se recomienda trabajar con una base de datos balanceada, sin embargo, se ha expuesto en la la tabla 1.1 que las bases de datos de imágenes faciales públicas más populares empleadas actualmente se caracterizan cada una por su propio sesgo (edad, color...), destacando en todas ellas una distribución no uniforme de las clases etiquetadas. Conscientes del problema que esto conlleva, se ha generado una base de datos de imágenes faciales, denominada DiveFace, considerada como un recurso muy valioso, debido a que la amplia diversidad de características de los sujetos favorece el rendimiento y la mejora de la biometría facial.

Además, en este trabajo se presenta una nueva formulación de discriminación algorítmica con aplicación al reconocimiento facial propuesta por el grupo BiDA-Lab. De acuerdo con la formulación propuesta, se han evaluado dos de los modelos más populares, VGG Face y ResNet-50. En primer lugar, se ha explorado hasta que punto los modelos VGGFace y ResNet-50 clasifican el género y la etnia por medio del entrenamiento de capas de clasificación conectadas a las capas de extracción de características de los modelos. Los resultados sugieren que estas redes son capaces de discriminar entre género y etnia sin que exista un gran deterioro del rendimiento a pesar de suprimir casi el 50 % de las neuronas de la capa de *embeddings*. VGG Face y ResNet-50 son modelos sesgados basados en algoritmos competitivos de aprendizaje profundo, que han demostrado alta sensibilidad a atributos como el género o la etnia. Esta sensibilidad principalmente se traduce en un rendimiento muy diferente en función del grupo étnico.

Por último, se ha realizado el entrenamiento sobre arquitecturas propias con el objetivo de comprender el funcionamiento o la influencia del sesgo en las mismas. Los resultados obtenidos fomentan el entrenamiento de modelos y métodos que abarquen la diversidad y las diferencias inherentes a cada grupo demográfico.

Como conclusión decir que el desarrollo de este trabajo, me ha llevado a considerar la gran importancia que tienen los atributos de género y etnia en la toma de decisiones de las tecnologías de reconocimiento facial más utilizadas. A pesar de que estos atributos son útiles para el reconocimiento facial, existe el riesgo de utilizarlos con fines no éticos.

El análisis final objeto de estudio de este trabajo, nos hace tomar conciencia de la magnitud del problema que puede suponer un algoritmo discriminatorio en sus diversas aplicaciones, y por ello nos conduce a generar modelos equitativos que garanticen la igualdad con independencia de aquellos atributos inherentes al ser humano como el género o la etnia.

5.2. Trabajo futuro

Como trabajo futuro, se presenta la mejora de la base de datos DiveFace, frente a posibles errores en la clasificación de las imágenes sobre los seis grupos demográficos presentes. Además, sería interesante ampliar la base de datos con imágenes que abarquen la más amplia pluralidad de grupos humanos en el mundo.

También, se propone estudiar la influencia de otras covariables como, por ejemplo, la edad u otros rasgos faciales en los sistemas de reconocimiento facial.

Por último, se plantea la posibilidad de encontrar en los algoritmos de reconocimiento facial los distintos factores en su diseño o implementación, que generan los resultados sesgados.

Bibliografía

[1] W. Knight, "Microsoft is creating an oracle for catching biased AI algorithms," *MIT Technology Review*, 2018.

[2] S. Barocas and A. D. Selbst, "Big data´s disparate impact," 2016.

[3] J. Buolamwini and T. Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," *Proc. of the ACM Conf. on Fairness, Accountability, and Transparency, New York, USA*, vol. 81, pp. 1–15, 2018.

[4] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large scale face recognition," *In Proc. of European Conf. on Computer Vision, Amsterdam, The Netherlands*, 2016.

[5] I. Kemelmacher-Shlizerman, S. Seitz, D. Miller, and E. Brossard, "The MegaFace Benchmark: 1 Million Faces for Recognition at Scale," *In Proc. of IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas, USA*, pp. 4873–4882, 2016.

[6] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising face across pose and age," *In Proc. of Int. Conf. on Automatic Face and Gesture Recognition, Xian, China*, pp. 67–74, 2018.

[7] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep Face Recognition," *In Proc. of British Machine Vision Conf., Swansea, UK*, 2015.

[8] L. Wolf, T. Hassner, and I. Maoz, "Face Recognition in Unconstrained Videos with Matched Background Similarity," *In Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, Colorado Springs, USA*,, pp. 529–534, 2011.

[9] S. Yi, Z. Lei, and S. Z. Li, "Learning face representation from scratch," *arXiv:1411.7923*, 2014.

[10] S. Yang, P. Luo, C. C. Loy, and X. Tang, "From Facial Parts Responses to Face Detection: A Deep Learning Approach," *In Prof of IEEE Int. Conf. on Computer Vision, Santiago, Chile*, pp. 3676–3684, 2015.

[11] N. Kumar, A. Berg, P. N. Belhumeur, and S. Nayar, "Describable visual attributes for face verification and image search," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, no. 10, pp. 1962–1977, 2011.

[12] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, and P. Grother, "IARPA Janus benchmark - C: Face dataset and protocol," *In Proc. of Intl. Conf. on Biometrics, Gold Coast, Australia*, 2018.

[13] Z. Zhang, Y. Song, and H. Qi, "Age progression/regression by conditional adversarial autoencoder," *In Proc. of Intl. Conf. on Computer Vision and Pattern Recognition, Honolulu, USA*, pp. 5810–5818, 2017.

[14] E. Learned-Miller, G. B. Huang, A. RoyChowdhury, H. Li, and G. Hua, "Labeled Faces in the Wild: A Survey. Advances in Face Detection and Facial Image Analysis," *Edited by Michal Kawulok, M. Emre Celebi, and Bogdan Smolka, Springer*, pp. 189–248, 2016.

[15] J. Ortega-Garcia, J. Fierrez, *et al.*, "The Multi-Scenario Multi-Environment BioSecure Multimodal Database (BMDB)," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 32, no. 6, pp. 1097–1111, 2010.

[16] A. Morales, J. Fierrez, and R. Vera-Rodriguez, "Sensitivenets: Learning Agnostic Representations with Application to Face Recognition," *arXiv:1902.00334*, 2019.

[17] K. Hao, "This is how AI bias really happens — and why it's so hard to fix," *Recuperado de: https://www.technologyreview.com/s/612876/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/*, 2019.

[18] A. Moreno, "Algorithmic Discrimination: Formulation and Exploration in Deep Learning-based Face Biometrics," 2019.

[19] U. Europea, "Reglamento (UE) 2016/679 del Parlamento europeo y del Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la directiva 95/46/CE (Reglamento general de protección de datos). Diario oficial de la Unión Europea, 27.," .

[20] "Programa electoral PSOE," pp. 226–227, 2019. Download.

[21] C. Garvie, A. Bedoya, and J. Frankle, "The Perpetual Line-Up: Unregulated Police Face Recognition in America. Georgetown Law, Center on Privacy Technology," 2016.

[22] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. V. Bruegge, and A. K. Jain, "Face recognition performance: Role of demographic information," *IEEE Trans. on Information, Forensics and Security*, vol. 7, no. 6, pp. 1789–1801, 2012.

[23] C. Cook, J. Howard, Y. Sirotin, J. Tipton, and A. Vemury., "Demographic Effects in Facial Recognition and Their Dependence on Image Acquisition: An Evaluation of Eleven Commercial Systems," *IEEE Trans. on Biometrics, Behavior, and Identity Science*, vol. 1, no. 1, pp. 32–41, 2019.

[24] J. Zhao, M. Y. T. Wang, V. Ordonez, and K. Chang, "Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints," *In Proc. of Conf. on Empirical Methods in Natural Language Processing*, pp. 2979–2989, 2017.

[25] I. Kemelmacher-Shlizerman, S. Seitz, D. Miller, and E. Brossard, "The MegaFace Benchmark: 1 Million Faces for Recognition at Scale," *In Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, Las Vegas, USA*, pp. 4873–4882, 2016.

[26] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borthand, and L.-J. Li, "The new data and new challenges in multimedia research," 2015.

[27] R. L. Guerra, "¿existen o no las razas humanas?," *NosOtros. Recuperado de: http://www.cubadebate.cu/noticias/2011/11/11/existen-o-no-las-razas-humanas/.XL0TmjAzbIU*, 2011.

[28] I. Raji and J. Buolamwini, "Actionable Auditing: Investigating the Impact of Publicly Naming Biased

Performance Results of Commercial AI Products," *In Proc. of Conf. on Artificial Intelligence, Ethics, and Society, Honolulu, USA*, 2019.

[29] T. Calders and S. Verwer, "Three Naive Bayes Approaches for Discrimination-Free Classification," *Data Mining and Knowledge Discovery*, vol. 21, no. 2, pp. 277–292, 2010.

[30] F. Alonso-Fernandez, J. Fierrez, and J. Ortega-Garcia, "Quality Measures in Biometric Systems," *IEEE Security Privacy*, vol. 10, no. 9, pp. 52–62, 2012.

[31] R. Ranjan, S. Sankaranarayanan, A. Bansal, N. Bodla, J.-C.Chen, V. Patel, C. Castillo, and R.Chellappa, "Deep Learning for Understanding Faces: Machines May Be Just as Good, or Better, than Humans," *IEEE Signal Processing Magazine*, no. 35, pp. 66–83, 2018.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *In Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition, Las Vegas, USA*, p. 770–778, 2016.


UAM
UNIVERSIDAD AUTONOMA
DE MADRID